DOCUMENT RESUME

ED 164 609 TM 008 177

AUTHOR Quinn, Jimmy L.
TITLE Estimation of the Regression Effect Using a Latent Trait Model.
PUB DATE 28 Mar 78
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Correlation; Intermediate Grades; *Item Analysis; Mathematical Models; Reading Tests; *Sampling; Standard Error of Measurement; *Statistical Bias; Test Reliability
IDENTIFIERS *Birnbaum 3 Parameter Logistic Model; *Regression

ABSTRACT

A logistic model was used to generate data to serve as a proxy for an immediate retest from item responses to a fourth grade standardized reading comprehension test of 45 items. Assuming that the actual test may be considered a pretest and the proxy data may be considered a retest, the effect of regression was investigated using a percentage of the group from the lowest tail of the pretest distribution. The mean of the retest group is assumed unbiased by errors of measurement. Thus the means of the pretest and the proxy test should differ, on the average, by an amount equal to the regression effect. Regression effects for various sizes of samples tested, from 25 to 500, and size cuts, from the fourth percentile to the 89th percentile, were reported. Item parameters had been estimated by Birnbaum's three-parameter logistic model separately for four groups of 1,000 individuals each, selected by sex and by race. The regression effect was most powerful with samples from the extreme lower tail, and decreased to a low, stable level at about the 30th or 40th percentile. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED164609

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ESTIMATION OF THE REGRESSION EFFECT USING A LATENT TRAIT MODEL

Presented to the
American Educational Research Association
at its 1978 Annual Meeting in
Toronto, Ontario, Canada

by

Jimmy L. Quinn
Graduate Student in Educational Research
College of Education
University of South Carolina
Columbia, South Carolina

March 28, 1978
Printed in USA

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jimmy L. Quinn

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM008 177

TABLE OF CONTENTS

	Page
Abstract	1
The Regression Effect	2
The Selection of a Test for Study	4
Latent Trait Theory	5
Methodology	7
Results	10
Conclusion	11
Acknowledgements	13

LIST OF TABLES

Table		Page
I	REPORT OF SUBJECTS IN STATEWIDE TESTING	14
II	NUMBER OF SUBJECT IN SUBSET FROM WHICH ESTIMATES WERE COMPUTED	15
III	CORRELATION OF PRETEST RAW SCORES AND RAW SCORES GENERATED FROM ESTIMATES	16
IV	NUMBER OF RAW SCORES AT EACH RAW SCORE LEVEL AND CUMULATIVE PERCENTAGES	17
V	OBSERVED REGRESSION EFFECT	18

LIST OF FIGURES

Figure		Page
1	PLOT OF OBSERVED AND GENERATED RAW SCORES FOR THE ESTIMATION SAMPLE OF WHITE BOYS	19
2	THE OBSERVED MEAN REGRESSION EFFECT ON SAMPLES OF VARIOUS SIZES SELECTED AT OR BELOW VARIOUS RAW SCORES	20
3	THE OBSERVED MEAN REGRESSION EFFECT ON SAMPLES OF VARIOUS SIZES SELECTED AT OR BELOW VARIOUS CUMULATIVE PERCENTAGES	21
4	STANDARD DEVIATION OF OBSERVED REGRESSION EFFECT FOR SAMPLES OF VARIOUS SIZES SELECTED FROM BELOW VARIOUS RAW SCORES	22
5	OBSERVED MEAN OF PRETEST SCORES AS A FUNCTION OF SAMPLE SIZE AND THE RAW SCORE AT OR BELOW WHICH THE SAMPLE WAS SELECTED	23
6	OBSERVED RETEST (POSTTEST) MEAN SCORE AS A FUNCTION OF SAMPLE SIZE AND THE RAW SCORE AT OR BELOW WHICH THE SAMPLE WAS SELECTED	24

Abstract

A logistic model is used to generate a proxy for an immediate retest from item responses to an actual test. Assuming that the actual test may be considered a pretest and the proxy a retest, a percentage of the group from the lowest tail of the pretest distribution is selected. The mean of the retest group is assumed unbiased by errors of measurement. Thus, the means of the two testings should differ, on the average, by an amount equal to the regression effect. Regression effects for various sizes of groups tested and for various sizes cut from these groups are reported.

The Regression Effect

The regression effect is a common source of bias in improperly selected experimental groups. The bias results from the fact that measurement error is a function of observed score. Assume a function in which the ordered pairs are (x,y), such that x is the observed score and y is the expected value of bias due to the regression effect. Over the domain of observed scores, y would be a maximum at the lowest raw score, and would be a minimum at the highest raw score. The graph of y on x would be monotonically decreasing. The graph of y would cross the x axis at the mean raw score. The source of the bias is readily found. The relationship between observed score (O) and true score (T) is

$$O = T + e$$

where e is measurement error. Measurement error is assumed to be distributed normally over all measurements. Observed score has utility only insofar as it mirrors true score. Consider a subject whose true score places him in the lower tail of the distribution of true scores. Observed score is determined. If the direction of error in his observed score is positive, such an error will move his observed score out of the extreme end of the distribution of observed scores. A subject with a true score slightly higher than his may have an observed score in the tail because of negative error. Thus, the extremely low observed scores are far more likely to contain negative error than positive error. An exactly opposite situation obtains in the upper tail, where positive error in observed scores outweighs negative error. The imbalance of measurement error in the tails of the distribution produces the regression effect.[1]

---

1. See, for example, Quentin Stodola and Kalmer Stordahl, Basic Education Tests and Measurement. Chicago: Science Research Associates, Incorporated, 1967. p. 282 Another excellent source is Statistical Theories of Mental Test Scores, by Frederick M. Lord and Melvin R. Novick, Reading,Mass.: Addison-Wesley Publishing Co., 1968

The presence of the regression effect in the tails causes concern: 1) when a subset of subjects is selected from one of the tails of the distribution, and 2) the mean of the scores of that subgroup is used as a pretest mean for later comparison. Clearly, that subset is relatively homogeneous with regard to the degree of their possession of whatever trait the test measures. Imagine that a subgroup from the lower end of the distribution of raw scores is selected by taking those subjects whose raw scores are less than k, where k is a raw score relatively far below the mean. The mean of these raw scores is quite likely to be too low. An improvement in mean raw score could be predicted if this group were tested again with a parallel test, even though no real improvement had occurred. Many evaluative studies have used the weak pretest-posttest design without consideration of the regression effect. For these studies, a sample was selected from the lower tail of scores. The mean of these scores was used as the pretest mean. When gains were reported, no credit was given to the regression effect, either because of ignorance or outright deception. Studies which contain such error continue to be conducted, partly because of the unwillingness of educators to "waste time in additional testing." They resent the additonal testing of the selected subgroup with a parallel form of the test; such "double-testing" is the respected procedure for avoiding large regression effects in experimental and evaluative designs.

An alternative solution to the problem of regression effect bias would be available if the regression of error on raw score were well-understood and predictable. In such a case, one would need to adjust the pretest mean by an appropriate amount so that the expected regression effect was countered. Once the size of the correction was well-known and evaluators realized that such a correction was to be made to their data, a predictable result would be that studies which

foster large regression effects probably would be redesigned.

The remainder of this paper details a method used by the writer to assess the magnitude of the regression effect.

The Selection of a Test for Study

The present study could be made without reference to actual test data simply by creating some data which fit some defensible model. The challenge of working with real data is appealing, however. Should real test data be used, a sufficiently large quantity of score vectors must be available on the basis of which item statistics could be estimated. Further, the test should be thoroughly researched and have a considerable body of normative information in its technical manual. Finally, if the results of this study are to be practical, the test should be one which is popular among evaluators who feel the need to minimize the time subjects spend in testing. A test which satisfies all these demands is Comprehensive Tests of Basic Skills, produced by the California Testing Bureau. In several years of use in South Carolina as the achievement battery for a near universal testing of fourth and seventh graders, the battery has produced a massive amount of data. The statewide use of the tests has made them familiar to evaluators and increased their popularity. The tests are used more than any other as a primary source of evaluative data for programs seeking to improve pupils' skills in elementary reading and arithmetic. Since the entire battery was too extensive for the time allotted for this work, a single subtest was selected. Thus for work reported here, the test which we shall call the "pretest" was the CTBS Reading Comprehension Test, Form S, Level 1, as it was written by S.C. fourth graders in the fall of 1974.[2]

---

2. Comprehensive Tests of Basic Skills Level 1 Form S Expanded Edition Examiner's Manual. Monterey, Calif.: CTB/McGraw-Hill, 1973.

Latent Trait Theory

As proposed by Allen Birnbaum, the probability that subject a will mark the answer to item i correctly on a test is given by:[3]

$$P_{ia} = \text{Prob}(u_{ia} = 1 \mid a_i, b_i, c_i, \Theta_a) =$$

$$c_i + (1 - c_i)\Psi(l_{ia}) \qquad \text{where}$$

$$l_{ia} = Da_i(\Theta_a - b_i),$$

and where

$$\Psi(l_{ia}) = \frac{1}{1 + \exp^{(-l_{ia})}}$$

D is a constant approximately equal to 1.7. Θ_a is the ability level of examinee a, a_i is the item discriminating power, b_i is the item difficulty and c_i is the lower asymptote of the item characteristic curve, sometimes called the "guessing parameter."

In the test under study, containing 45 items, estimation of 135 (45 x 3) parameters for the items themselves and as many unique estimates of ability as there are unique combinations of item responses provides many degrees of freedom for fitting the test data to the model. Even when the ability estimates are reduced dramatically to 46, one for each possible raw score (0 to 45), freedom to fit the data to the model is considerable. Confidence seems justified that such a model must capture in its estimates the primary sources of variation within a set of item responses. The process of marking an item correctly is considered random, but the model seems

---

3. Lord and Novick, op. cit., p. 400f

sufficient to estimate the enduring characteristics of the data; to presume that the random processes left free to vary may be considered to vary normally seems no leap of faith. In order, however, to permit an even more intimate adjustment of the model to the data, separate estimates of each item parameter and each ability estimate were made for each of four groups which experience has shown to vary in test performance. These groups are formed by subsetting the tested subjects by race and by sex. Left in the data set are two levels of sex (boys and girls) and two levels of race (white and black). Too few members of another race were tested for their inclusion in the analysis.

Hopefully, the many estimates which are mentioned here will reflect the unique character of the interaction of these subjects with this test. Should one use these estimates to create a new set of item vectors with the same enduring qualities, one would expect these new vectors to be a reasonable proxy for the result of an identically parallel test. This second test would be unaffected by retesting effect, a primary source of bias in test-retest investigations. The correlation between the original raw scores and the generated raw scores is expected to be imperfect; random processes still operate. The correlation should be at least as high as test-retest reliabilities reported by the test publisher, since the biasing effect of retesting has been avoided in the former. In essence, the generated set of item responses represents an immediate retest with the same instrument made by the same subjects from the brains of which all recollection of the earlier testing has been erased.

Methodology

For the work reported here, the computer program LOGIST[4] was used to estimate the parameters on each of four groups of subjects, broken down by sex and race. These subjects wrote the 45-item reading comprehension test, from the Comprehensive Tests of Basic Skills, S/1, part of the statewide testing program sponsored by the South Carolina Department of Education. All subjects were fourth-graders. From each subgroup, 1000 subjects were selected in a systematic sample from the statewide alphabetical listing by schools by school districts. LOGIST eliminated some observations because of too low scores (well below chance expectancy) and perfect scores. Thus the estimates were based upon somewhat fewer than the 1000 sampled subjects. Table II reports the actual number of pupils in each group on which the estimates were based.

The computer program produced 45 estimates each for item discrimination, item difficulty, and lower asymptote for each of the four groups (a total of 720 estimates for the items). Also produced were nearly 4000 estimates of ability for the nearly 4000 subjects in the four groups. Since the current work required the use of these estimates to create a new set of item response vectors, the great number of ability estimates was cumbersome. Use of these estimates would have required estimation of ability for every unique score vector in the statewide set, a total of 45,593. To shorten the task, the mean ability estimate at each level of raw score was computed within each subgroup and used throughout the investigation. Thus 172 estimates of ability were computed, since raw scores of 0, 1 and 45 were absent from the estimation of ability in LOGIST. For purposes of regeneration of the score vectors, a score of 0 was regenerated as a 0. A score of 45 was regenerated as 45. A score of 1 was assigned an

---

4. R. L. Wood, Marilyn Wingersky and Frederick M. Lord, "LOGIST; a Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters, Research Memorandum." Princeton, N.J., June, 1976.

ability estimate which was extrapolated by extending the graph of ability estimates at raw scores near 1. Thus fortified with 720 item parameters and (46 x 4) 184 ability estimates, a new set of score vectors was produced by electronic means. The probability function presented earlier is a cumulative distribution function which returns the probability that the item will be marked correctly. Using the RANDU subroutine from the Scientific Subroutines Package, a uniformly distributed number, free to vary randomly between 0 and 1, was produced and compared with the probability predicted by the formula. If the number from RANDU was equal to or less than the probability of a correct response, the subject was given a correct response. Otherwise, he "missed" the item. Raw scores for each artificial score vector were computed.

If one grants the prudence of considering the artificial set of vectors a reasonable substitute for the results of an immediate retest after washing the subjects' brains, the subsequent work reported here may seem appropriate. Without such faith in the artificial vectors, the rest of the report is irrelevant. Thus, the report will proceed from this point by calling the artificial response matrix the "retest" dataset.

Scatter diagrams of raw score in the pretest by raw score in the retest were examined in order to check on the procedure for recreating the vectors, and particularly to examine the inflation of "unexplained" variance in the retest scores which can be expected to accrue from the practice of using mean estimates of ability. (The scatter diagram of ability estimates on raw score revealed a high linear association. The conditional variance of ability at each raw score level was a decreasing function of raw score. The larger variance of ability at low raw scores is expected: subjects may mark a few easy items correctly, a few hard items correctly, or a mixture of

a few easy and hard items correctly, thus supporting widely different predictions of their separate abilities.) The pretest-retest scatter diagram for white boys is presented in Figure 1 as representative of the four scatter diagrams. A report of the correlation between the pretest and retest scores appears in Table III. The table and the scattergram show the recreation of scores to be performing in accord with expectations and vindicates, for our purpose, the aggregation of thetas at each raw score level.

A frequency table of raw scores was prepared from a systematic sample of 4000. Because of an abnormally high number of zero scores, all such scores were dropped from the analysis. Such exclusion is defended on the basis that a subject with a raw score of 0 may by regarded as untested. Samples were selected from the lower tail of the distribution of pretest scores. A certain raw score was established as a cut point so that that score and all lower scores were eligible for inclusion in the sample. The scores used as cut points in successive samplings were 7, 9, 11, 14, 18, 23, 28, 32, 36, and 40. These cut off scores selected samples which represented, first, 5% of the prestest distribution, and then 10%, 20%, . . . , 90%.

After selection of the cut scores for the samples, a systematic sample of 4000 pretest-retest raw score pairs was read into the computer. A program drew simple random samples from these paired observations, discarding each pair in which the pretest raw score exceeded the cut score. When 25 observations were selected below the cutoff, pretest and retest means were computed. Other paired observations then were added randomly to the set until 50 were accumulated. Two more means were computed. This process was continued until means were computed for 100, 250, and 500 observations in the sample, as well. At that point, the next higher cutoff score

was selected and the entire process was repeated, cycling as more and more of the lower tail of the distribution entered the sample. At the end of the process, another systematic sample was introduced to the computer and the entire process was repeated until 20 samples of each size were drawn at or below every cutpoint.

Results

Table IV reveals the nature of the distribution of raw scores in the population under study. The frequencies at various raw scores reveal two relative maxima of the distribution, near raw scores of 11 (chance expectancy is 11.25) and 41. The leftmost mode is the highest in the distribution, indicating a sizable portion of pupils whose performance is no better than one would expect if they marked their answers randomly. The rightmost mode, less impressive in profile, is more platykurtic and represents many more pupils with strikingly better performance.

Table V and Figure 2 reveal the observed difference between retest mean and pretest mean, which is assumed to be caused by the regression effect. When large samples (N=500) are selected at or below a raw score of 7 (5%, approximately), the regression effect biases the pretest mean by 3.26 raw score points. The bias decreases to 0.35 raw score points at or below a raw score of 18 (approximately 40%). For cuts which select more than 40% from the lower end of the distribution, the regression effect seems to stabilize near 0.45 raw score points. Figure 4 shows the standard deviations of the regression effects observed in the samples at various cut points. These standard deviations show considerable variability among the smaller samples. The central limit phenomenon serves to stabilize the

standard deviation about a well-defined axis for samples of 500. The axis appears to be horizontal. Thus, a reasonable approach to reporting the data graphed in Figure 4 is to report the mean of the observed standard deviations at each sample size. Such a report comprises Table VI.

Figure 3 is another plot of the data in Figure 2, with the horizontal axis scaled according to the percentile rank of the cut score instead of raw cut score alone.

As information, two additional figures are presented without extensive comment. Figure 5 reports the observed pretest mean of the sample means as samples were cut from the distribution at or below various raw scores. Figure 6 reports the same information, except that the retest means take the place of the pretest means. These two figures, taken together, reveal the change in mean scores at various cut points as a result of the regression effect.

Conclusion

A proxy for a retext matrix of score vectors was recreated within an electronic computer from estimates made upon the pretest dataset using the three parameter logistic model as presented by Allen Birnbaum. The regenerated raw scores correlated highly with the original raw scores ($r = .97$). The regression effect was studied at locations of interest in the lower tail of the distribution of pretest scores. The effect was estimated, tabulated and plotted in two metrics.

The regression effect is shown powerfully to bias the mean of samples drawn from the extreme lower tail of a distribution. In the population studied, at 5%, the bias is roughly half the cut score and is more than half of the mean itself. Compare the bias of 3.26 to the pretest mean in the 5% cut, which is 5.89. Had the bias been compensated, the pretest mean would have risen sharply to 9.15, a figure considerably greater than the cutoff raw score of 7. Though vitally important as a source of bias for samples drawn below the 20th percentile, the regression effect decreases sharply to a low, stable level at about the 30th or 40th percentile.

Acknowledgements

The writer wishes to thank Professor Garrett K. Mandeville, College of Education, University of South Carolina, for his cheerfully shared profound knowledge, timely advice, long-suffering patience and inspiration.

Statewide testing data for this project were supplied to the writer from files of the S.C. Department of Education by State Superintendent of Education Cyril B. Busbee. Retiring next year, Dr. Busbee has brought a vision to his office which has led to the creation of superb educational data resources for many years to come. Dr. Paul M. Sandifer, Director of the Department's Office of Research recommended the writer's request for data to Dr. Busbee. Dr. John Finch of the Office of Research provided magnetic tapes of score vectors and format information for their utilization.

Computer service was provided on the IBM/370 computer at the University of South Carolina by the Academic Services Section of the University Computer Services Division. Charles Poole, III, of that section was invaluable through his suggestions and direct computational assistance.

Though modified extensively, the graphic presentations as well as the statistics in this paper were generated by the Statistical Analysis System (SAS), a useful package of easily accessible statistical procedures which provide considerable computational simplicity for the user. Information about SAS is available in the manual by Anthony J. Barr, James H. Goodnight, et al., A User's Guide to SAS 76, (Raleigh, N.C.: SAS Institute, Inc., 1976).

Such excellent support as mentioned gratefully here would have made this project far more successful but for the limitations of the writer who claims all mistakes, albeit apologetically.

TABLE I

REPORT OF SUBJECTS IN STATEWIDE TESTING

		Race		
		White (N)	Black (N)	Total (N)
SEX	Boys (N)	13,405	9,396	22,801
	Girls (N)	13,102	9,690	22,792
	Total (N)	26,507	19,086	45,593

TABLE II

NUMBER OF SUBJECTS IN SUBSETS FROM WHICH ESTIMATES WERE COMPUTED

Subset of Population	Subjects Sampled (N)	Subjects Used for Estimation (N)
White Boys	1000	983
Black Boys	1000	971
White Girls	1000	961
Black Girls	1000	985

TABLE III

CORRELATION OF PRETEST RAW SCORES AND RAW SCORES GENERATED FROM ESTIMATES

Subset	Correlation Coefficient Pretest-Retest r	Prediction Equation Retest = (Pretest x b) + a	
		b	a
White Boys	0.973	0.984	0.702
Black Boys	0.952	0.987	0.641
White Girls	0.987	0.984	0.681
Black Girls	0.957	0.982	0.653

TABLE IV

NUMBER OF RAW SCORES AT EACH RAW SCORE LEVEL AND CUMULATIVE PERCENTAGES

Raw Score	Frequency	Cumulative Percentage	Raw Score	Frequency	Cumulative Percentage
0	deleted	.	23	79	50.06
1	0	.	24	80	51.85
2	0	.	25	67	53.36
3	5	0.11	26	82	55.20
4	12	0.38	27	95	57.33
5	29	1.03	28	89	59.33
6	45	2.04	29	83	61.19
7	80	3.84	30	87	63.15
8	113	6.38	31	109	65.60
9	154	9.84	32	95	67.73
10	157	13.36	33	112	70.24
11	199	17.83	34	99	72.47
12	177	21.81	35	121	75.19
13	178	25.80	36	118	77.83
14	150	29.17	37	127	80.69
15	130	32.09	38	120	83.38
16	121	34.81	39	118	86.03
17	111	37.28	40	122	88.77
18	108	39.73	41	130	91.69
19	90	41.75	42	126	94.52
20	108	44.17	43	108	96.95
21	93	46.26	44	75	98.63
22	90	48.28	45	61	100.00

TABLE V

OBSERVED REGRESSION EFFECTS

(Each estimate is reported in raw scores and is based upon 20 samples.)

	Number in Each Sample				
Cutoff Raw Score	25	50	100	250	500
7 (4 %ile)	3.354	3.299	3.193	3.227	3.262
9 (10 %ile)	2.070	2.048	2.105	2.110	2.107
11 (18 %ile)	1.152	1.260	1.235	1.206	1.222
14 (30 %ile)	0.598	0.523	0.491	0.427	0.456
18 (40 %ile)	0.532	0.338	0.317	0.360	0.340
23 (50 %ile)	0.438	0.444	0.303	0.328	0.375
28 (59 %ile)	0.354	0.437	0.510	0.498	0.475
32 (70 %ile)	0.606	0.586	0.534	0.429	0.441
36 (78 %ile)	0.302	0.293	0.453	0.430	0.406
40 (89 %ile)	0.482	0.453	0.443	0.452	0.435

TABLE VI

STANDARD DEVIATION OF OBSERVED REGRESSION EFFECT
AS A FUNCTION OF SAMPLE N

Sample N	Mean Standard Deviation of Regression Effect Raw Score Units
25	0.623
50	0.431
100	0.313
250	0.188
500	0.141

Figure 1

PLOT OF OBSERVED AND GENERATED RAW SCORES
FOR THE ESTIMATION SAMPLE OF WHITE BOYS

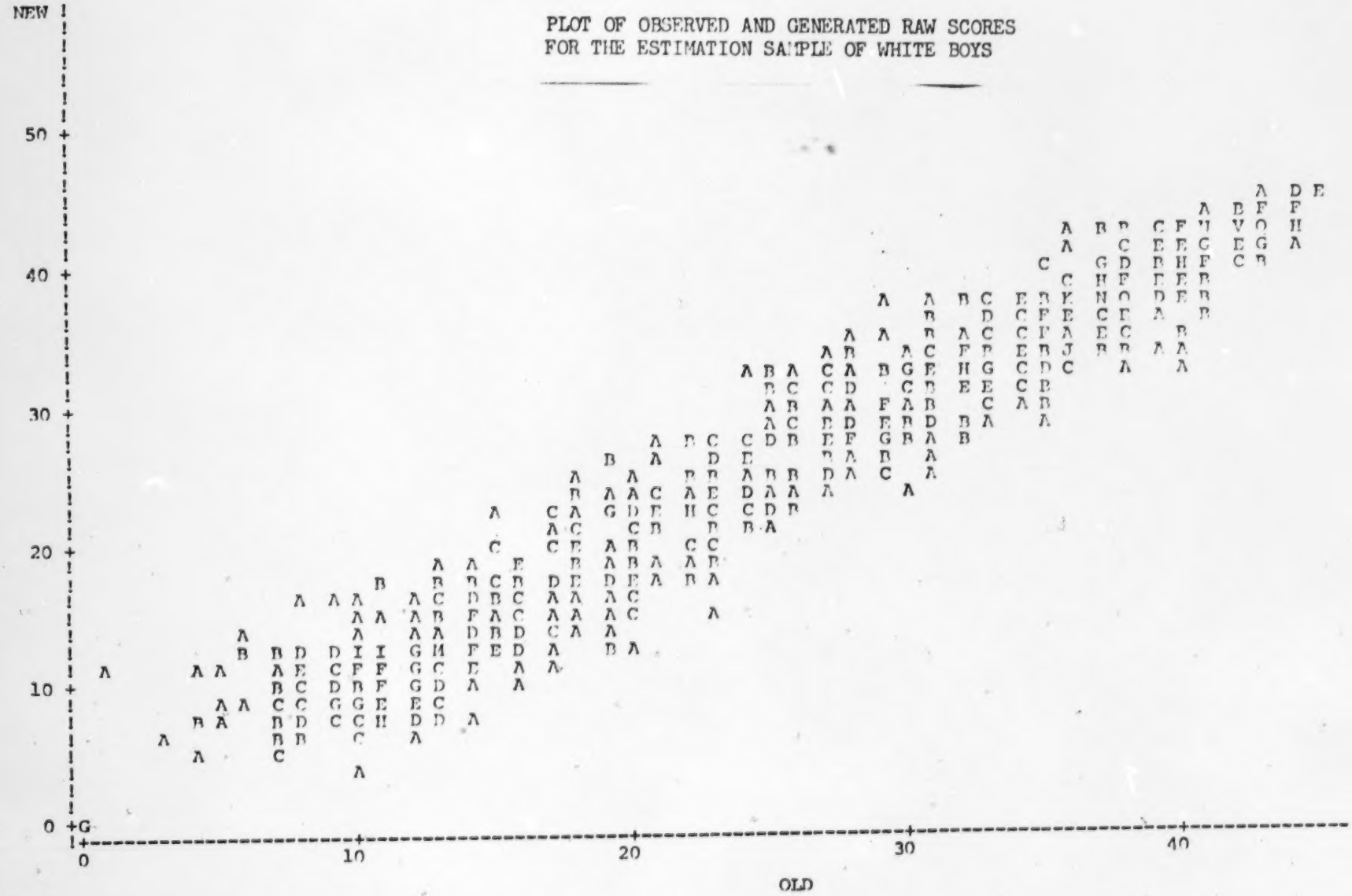

Figure 2

THE OBSERVED MEAN REGRESSION EFFECT ON SAMPLES OF VARIOUS SIZES SELECTED AT OR BELOW VARIOUS RAW SCORES

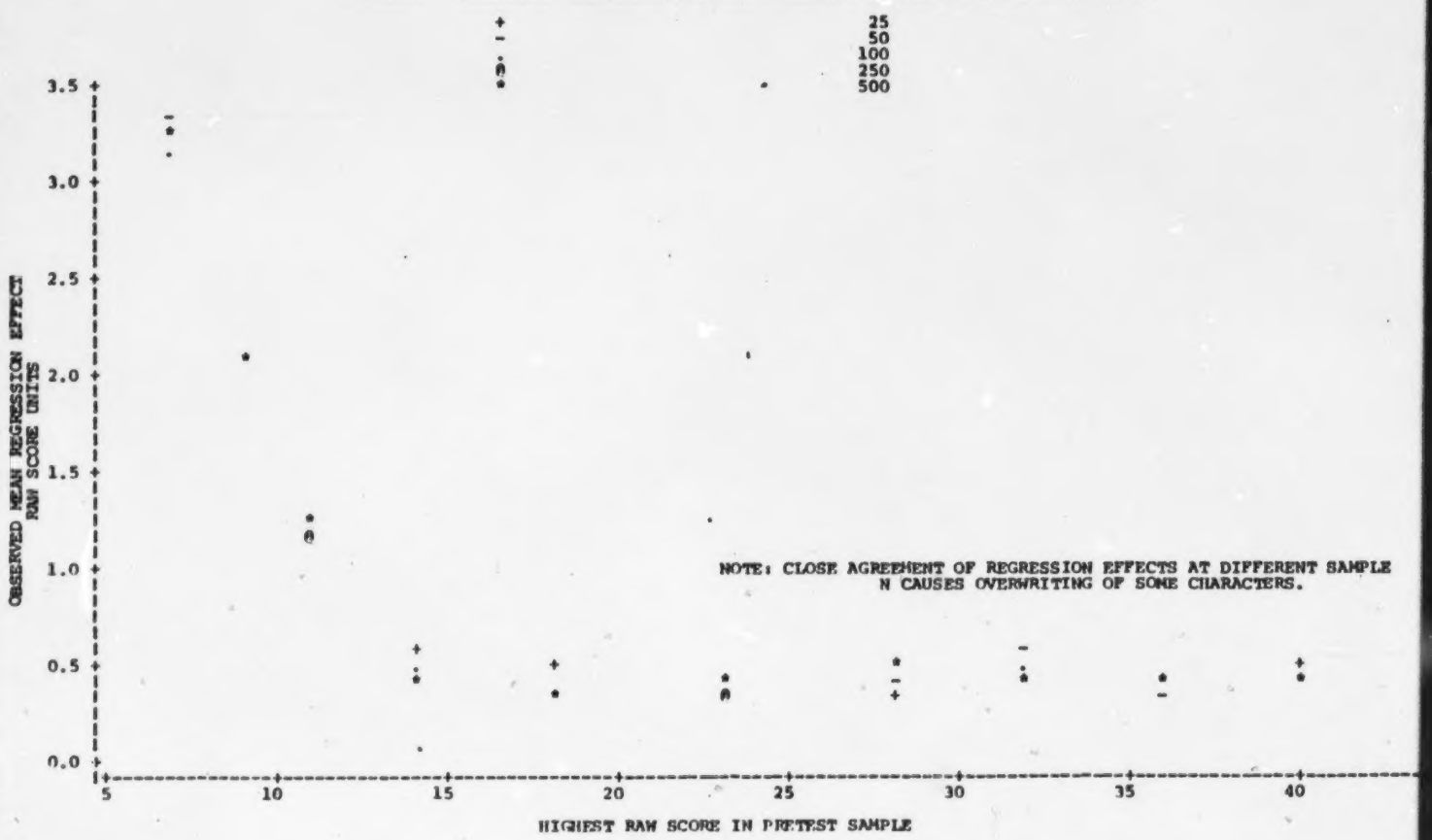

Figure 3

THE OBSERVED MEAN REGRESSION EFFECT ON SAMPLES OF VARIOUS SIZES
SELECTED AT OR BELOW VARIOUS CUMULATIVE PERCENTAGES

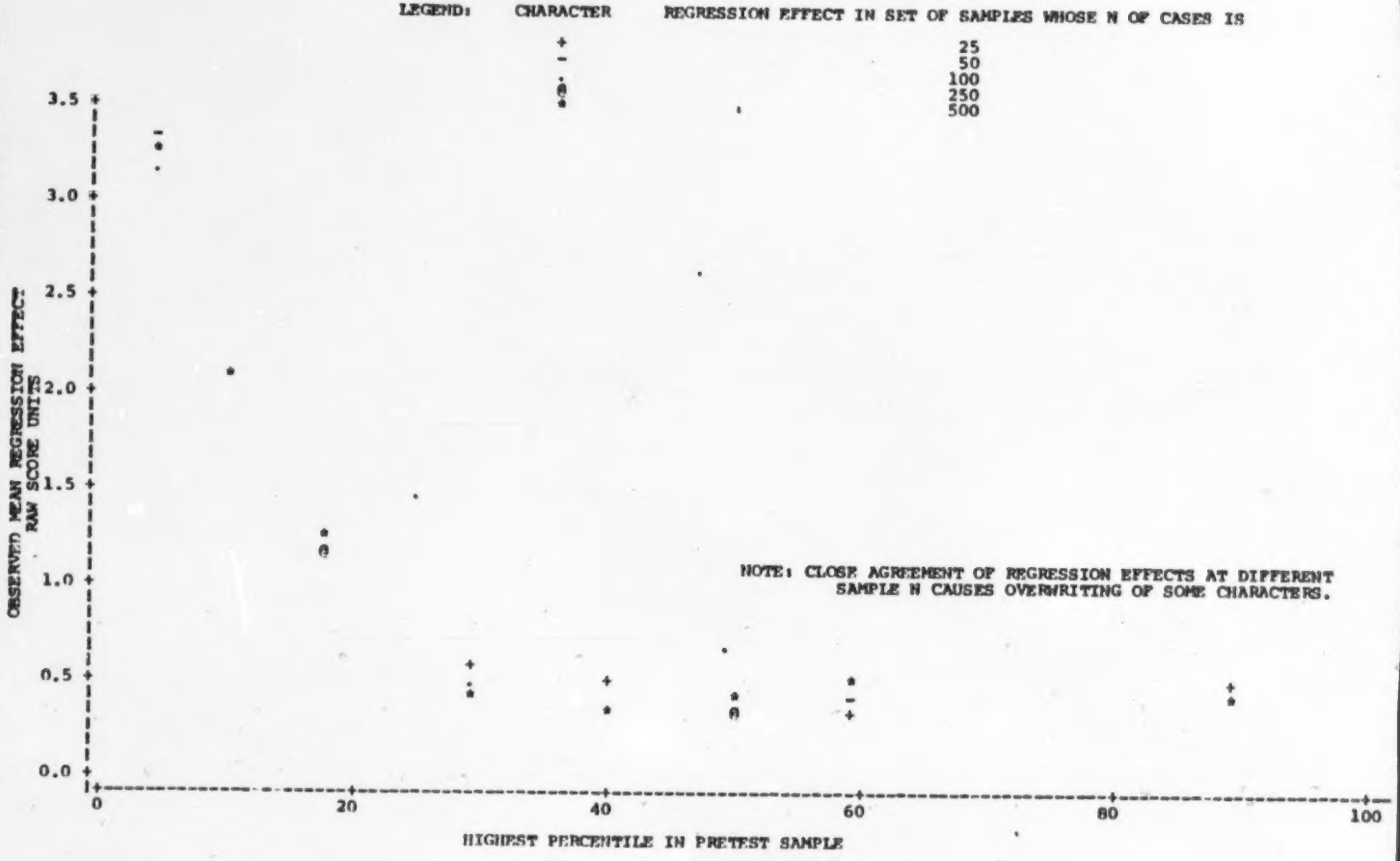

Figure 4

STANDARD DEVIATION OF OBSERVED REGRESSION EFFECT FOR SAMPLES OF VARIOUS SIZES SELECTED FROM BELOW VARIOUS RAW SCORES

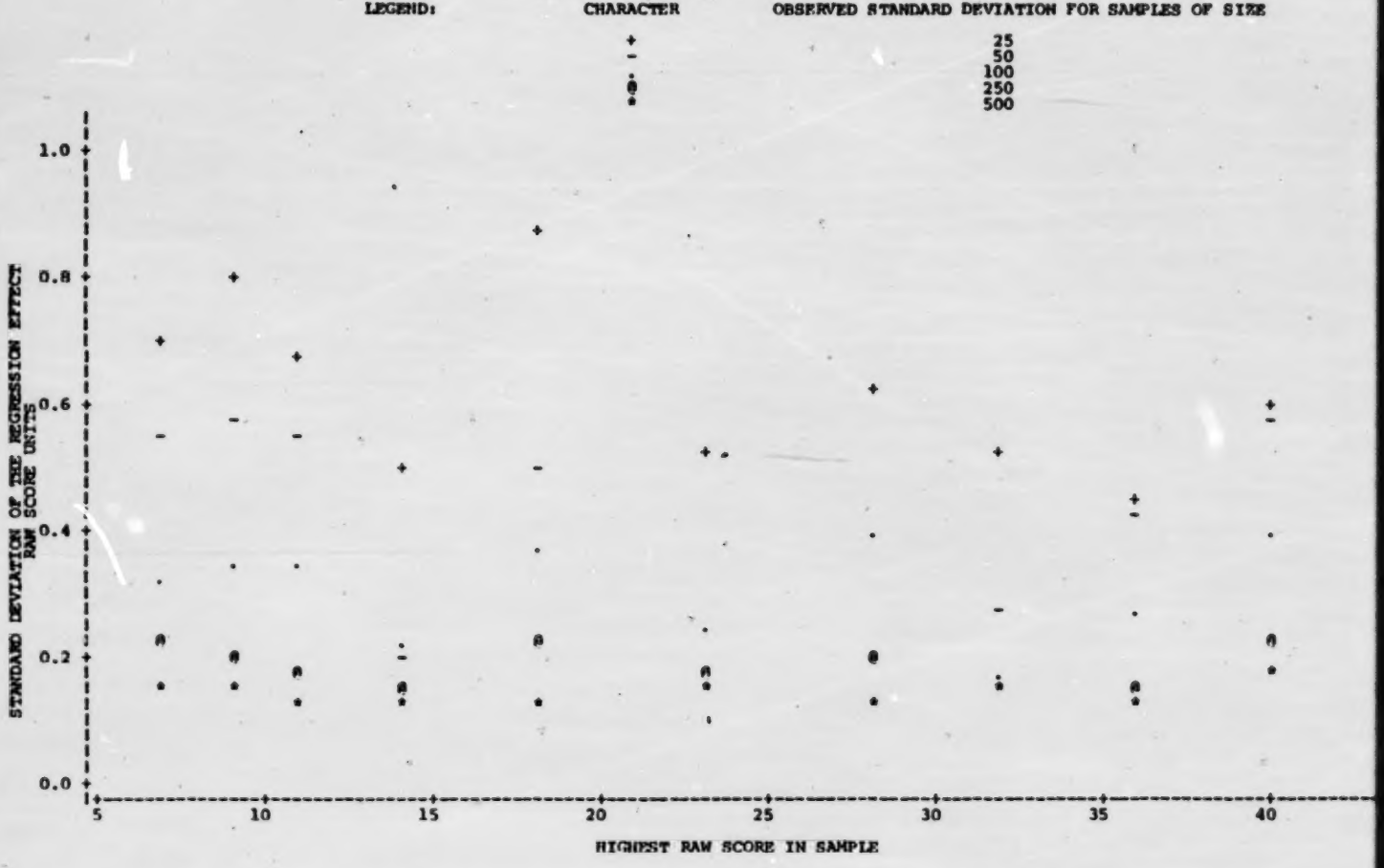

Figure 5

OBSERVED MEAN OF PRETEST SCORES AS A FUNCTION OF SAMPLE SIZE AND THE RAW SCORE AT OR BELOW WHICH THE SAMPLE WAS SELECTED

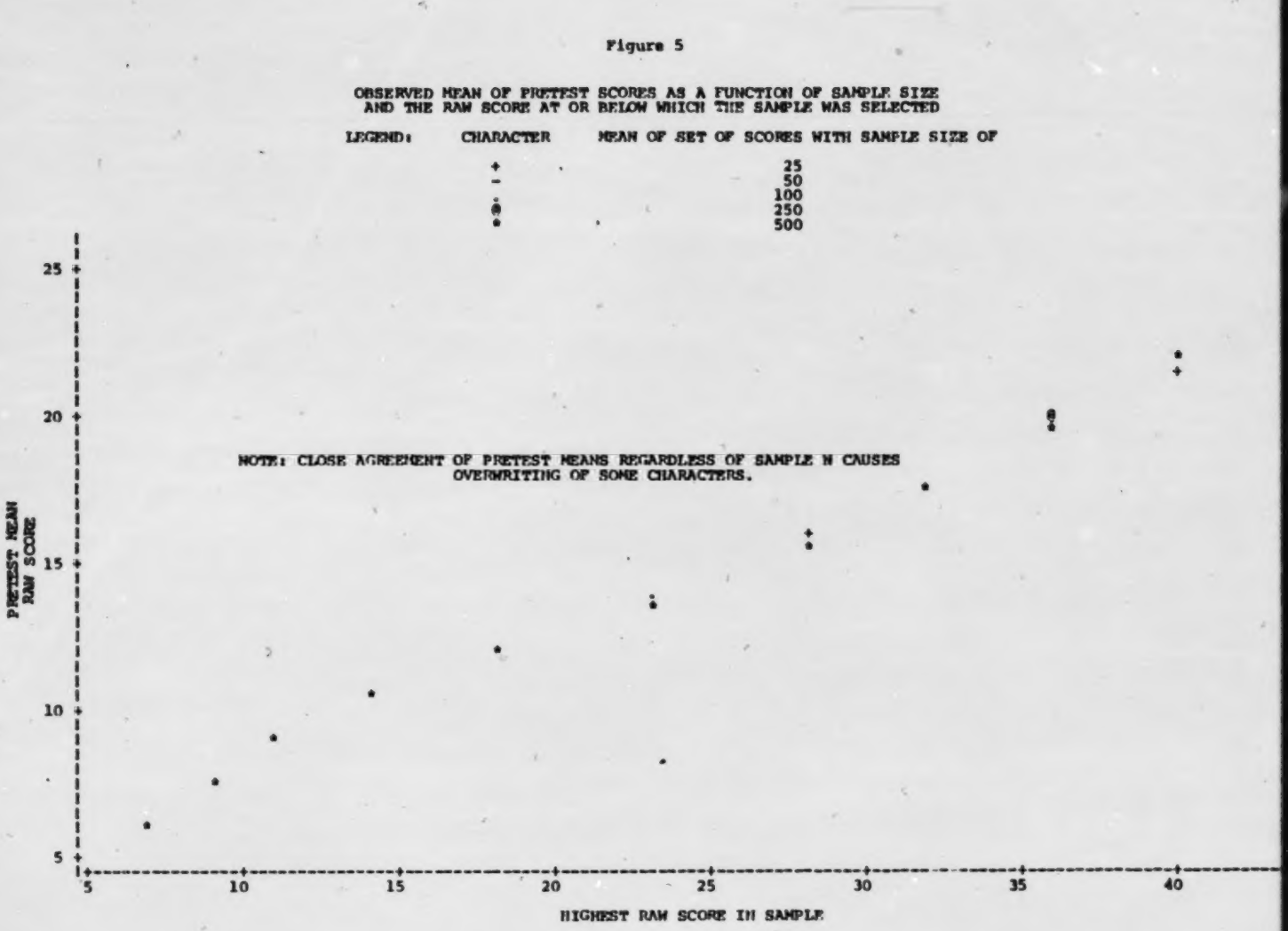

Figure 6

OBSERVED RETEST (POSTTEST) MEAN SCORE AS A FUNCTION OF SAMPLE SIZE AND THE RAW SCORE AT OR BELOW WHICH THE SAMPLE WAS SELECTED